

# Systems Level Analysis of Cancer Heterogeneity

Teresa Przytycka  
NIH / NLM / NCBI



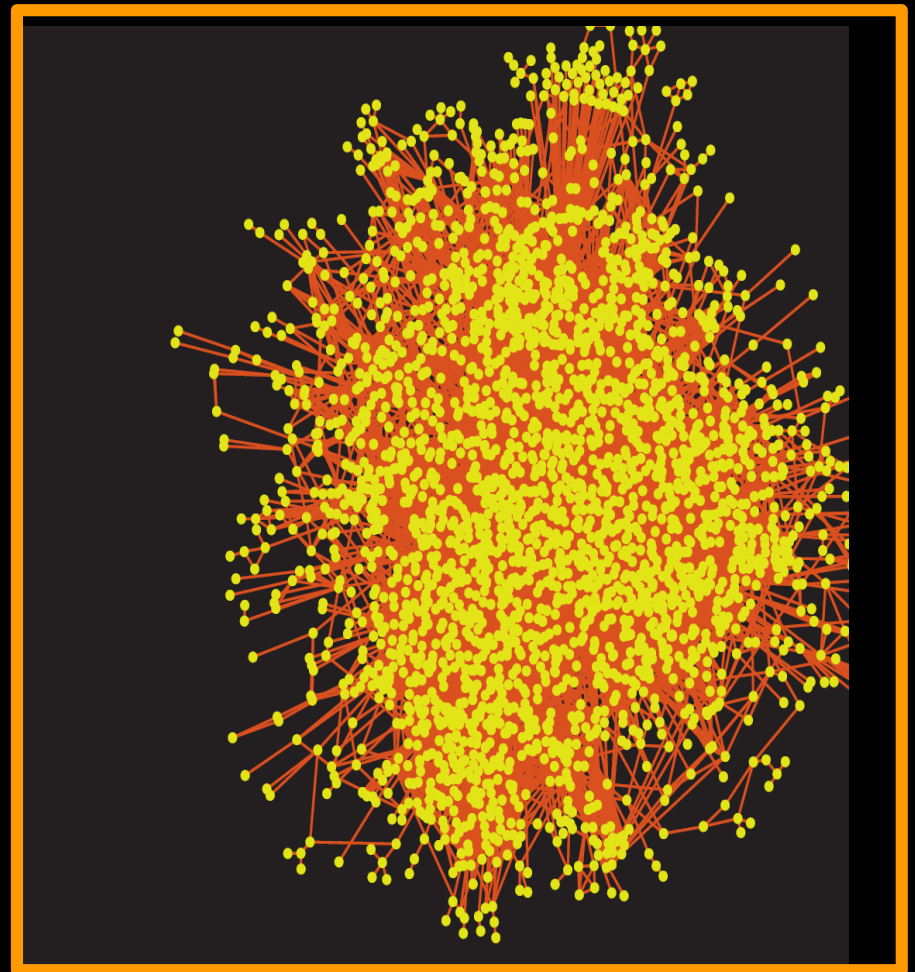
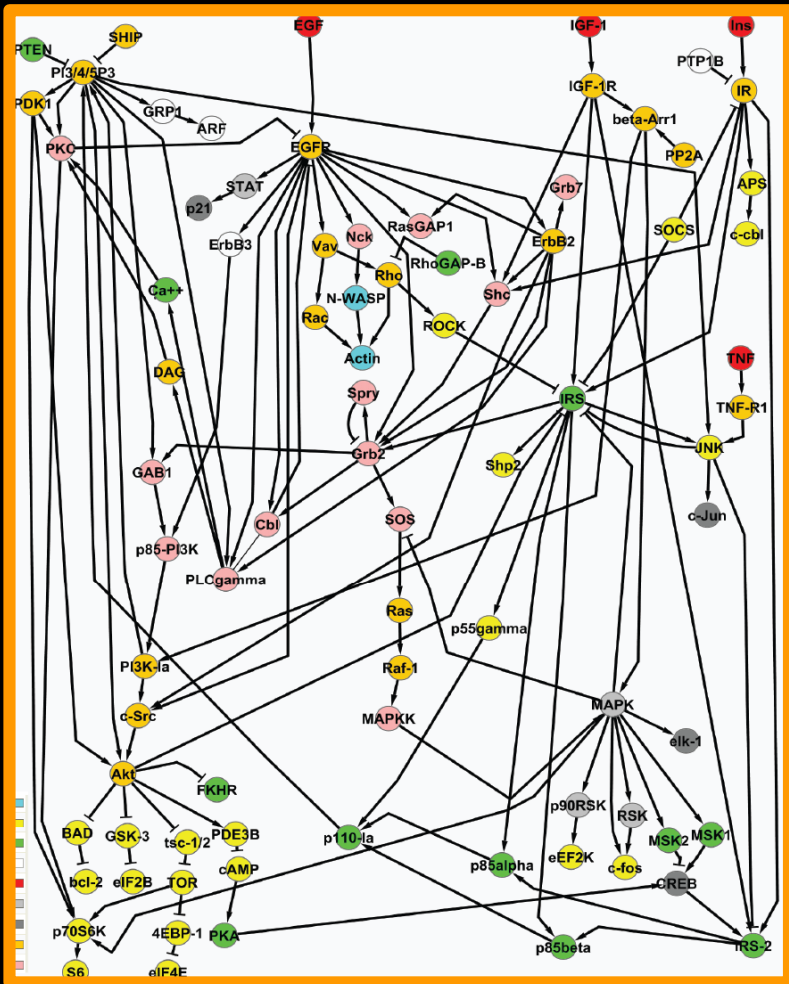
# Key challenges in cancer data analysis

- **Complexity:** Multiple driver mutations are typically required for cancer progression
- **Heterogeneity:** Phenotypically similar cancer cases might be caused by different sets of driver mutations
  - **Driver mutations /alterations**— mutations contributing to cancer progression
  - **Passenger mutations** — neutral mutations accumulating during cancer progression
- Some driver mutations are rare
- Epistasis — masking of the effect of one mutation by another mutation
- Cancer evolution

# Network/Systems biology view

- **Motivation:**
  - Effects of genetic alteration propagate through the network affecting downstream genes
  - Different driver mutations often dysregulate common pathways
- **Main lines of attack:**
  - Examining known pathways for a signature of dysregulation
  - Computational pathways discovery from high-throughput interaction data

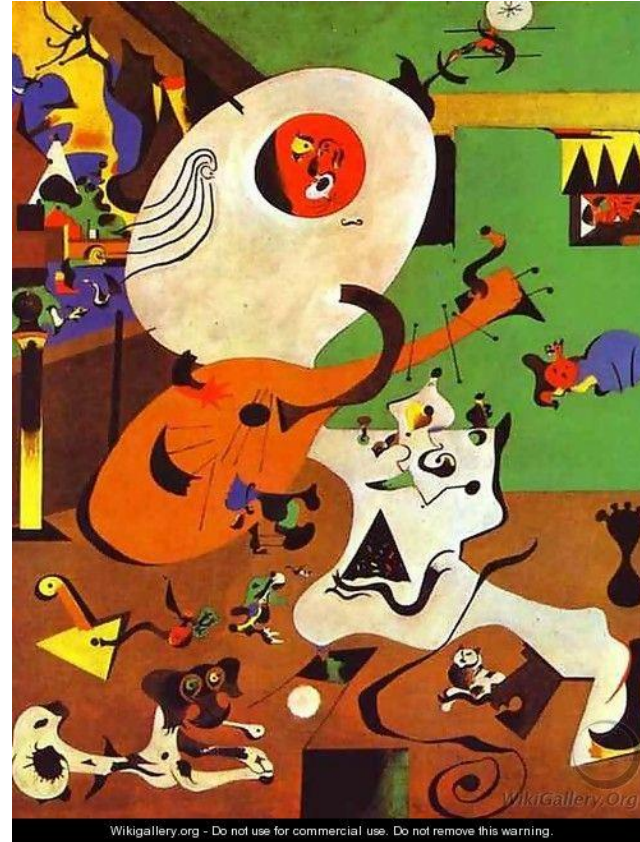
# Which network to use?



# High throughput network versus “the true” network



The Lute Player, Hendrick Maertensz Sorgh (1610-1670),  
Rijksmuseum, Amsterdam  
(public domain)



Dutch Interior 1, Joan Miró (1893-1983)  
Museum of Modern Art, New York  
© 2012 Successió Miró i Artista Rights Society (ARS), New York / ADAGP, Paris  
(used with ARS permission)

## Three general techniques that utilize network based approaches in cancer studies

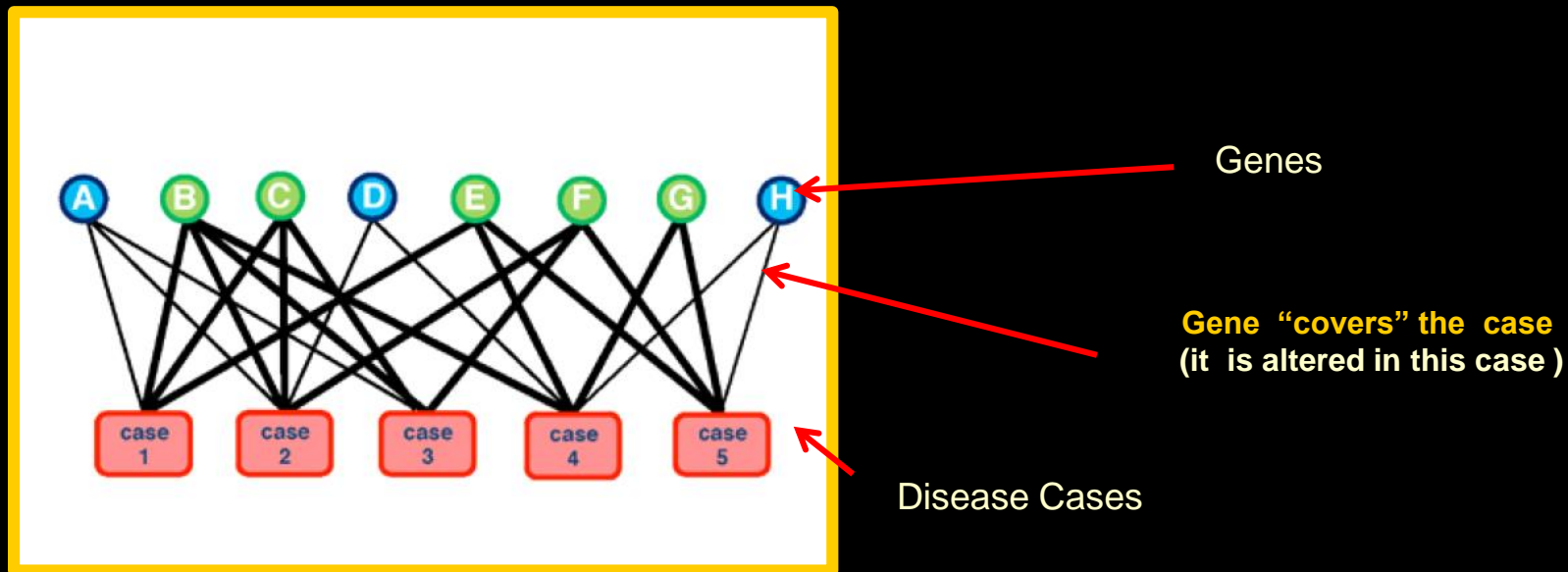
- Module cover
- Network Flow
- Mixture /topic models

## Three general techniques that utilize network based approaches in cancer studies

- Module cover
- Network Flow
- Mixture models

# Finding a representative set of dysregulated genes in disease cases

Goal: Given a set of dysregulated genes and disease cases, find a representative set of dysregulated genes



# Module Cover Approach

## Optimization problem:

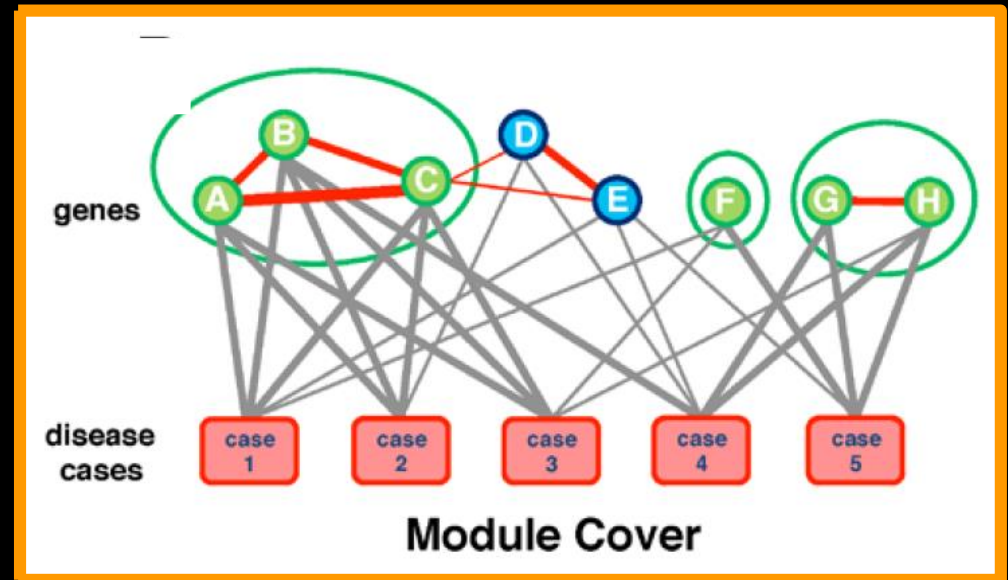
Find smallest cost set of modules so that each disease case is covered at least  $k$  times

**Cost** is a function of:

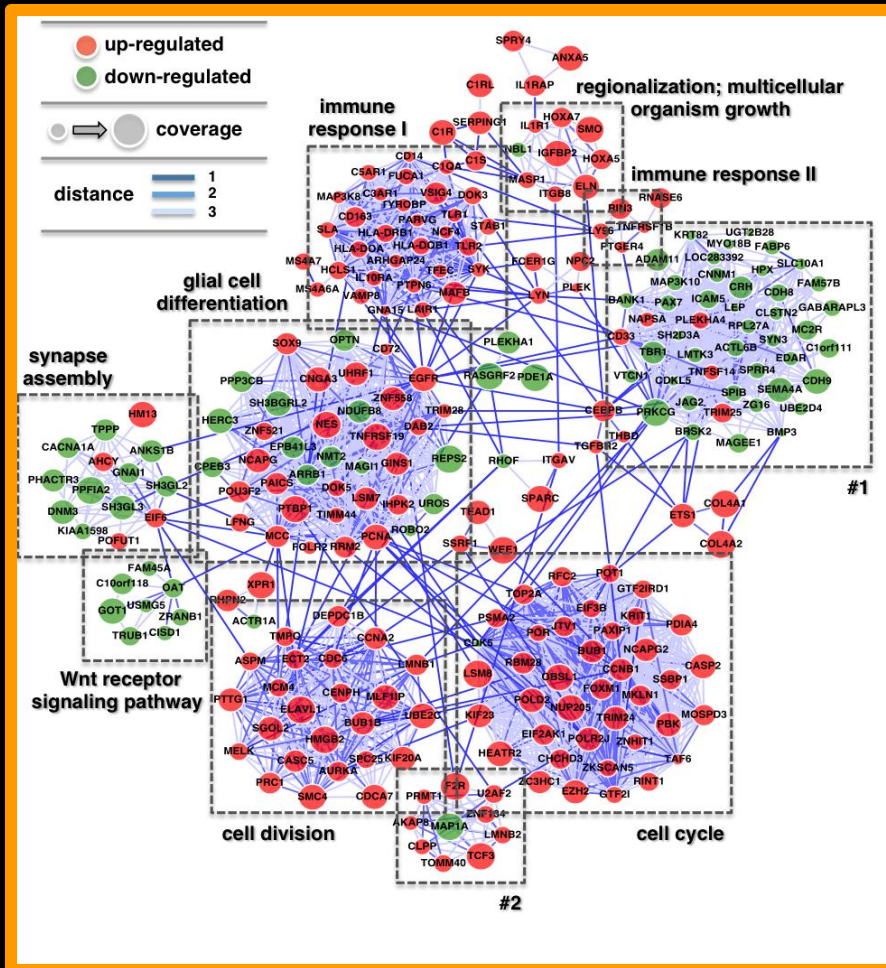
↓ distance in the network of genes in same module

↓ A similarity measure (application dependent)

↑ number of modules (parameterized penalty)



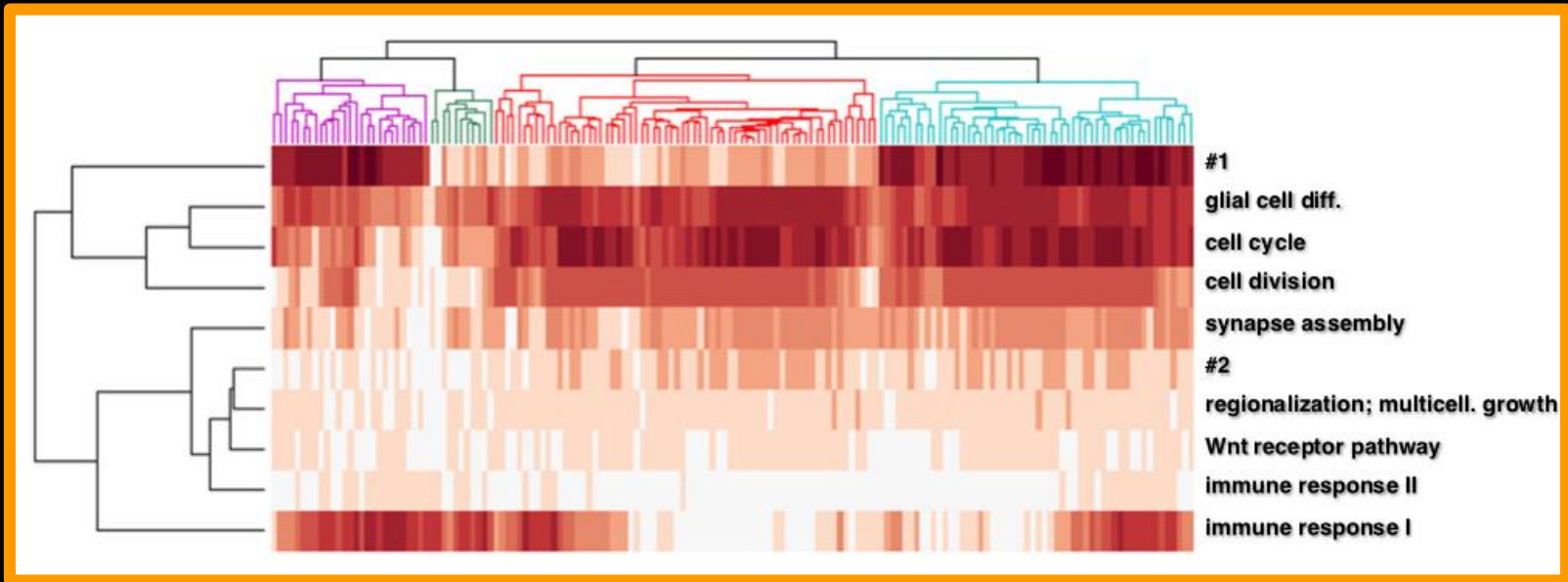
# Module Cover: Glioblastoma Data



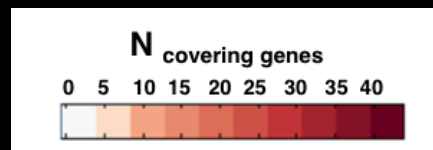
**Signature modules  
from GBM Dataset  
(REMBRANDT)**

# Different patients groups have different signature modules

cases



modules

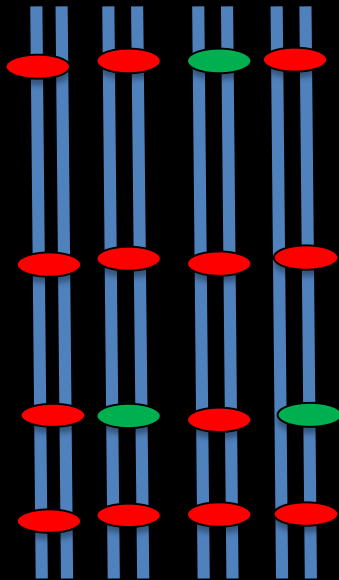


## Three general techniques that utilize network based approaches in cancer studies

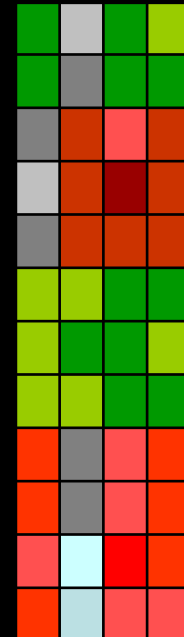
- Module cover
- Network Flow
- Mixture models

# Information flow from genotypic changes to expression changes

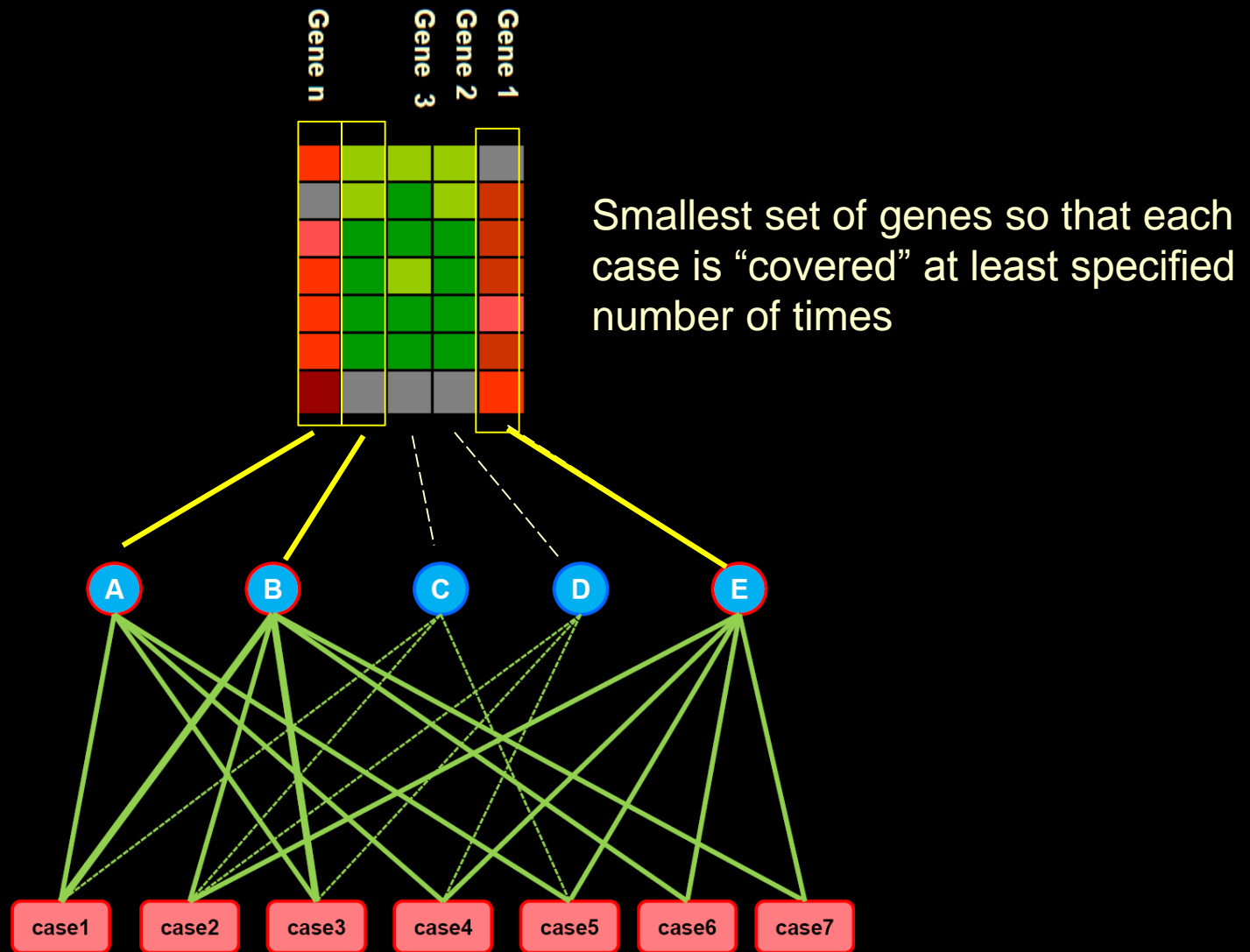
Copy number aberrations  
or/and mutations



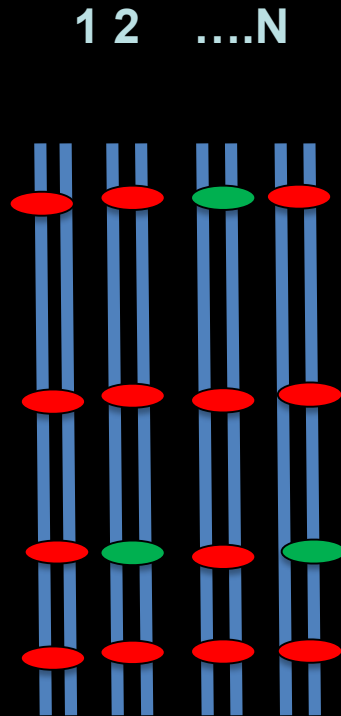
Gene expression



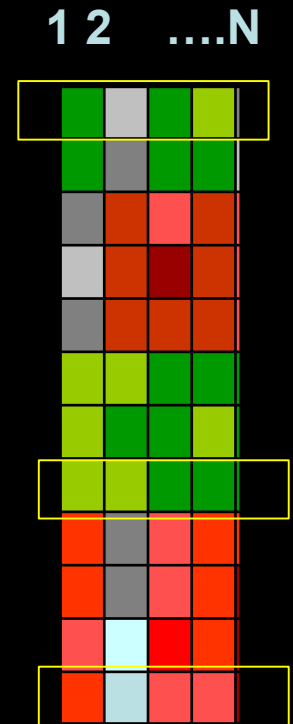
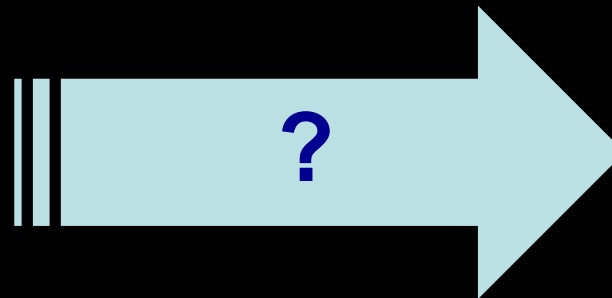
# Selecting “signature” genes



# Explaining expression changes in the signature genes

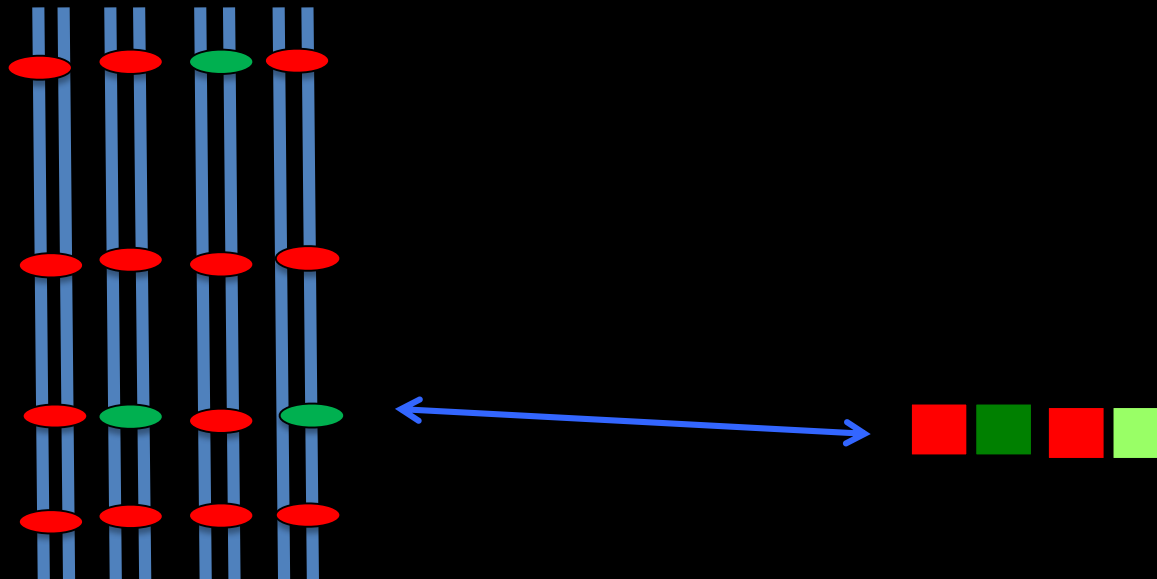


**Cancer Cases  
CNV data**



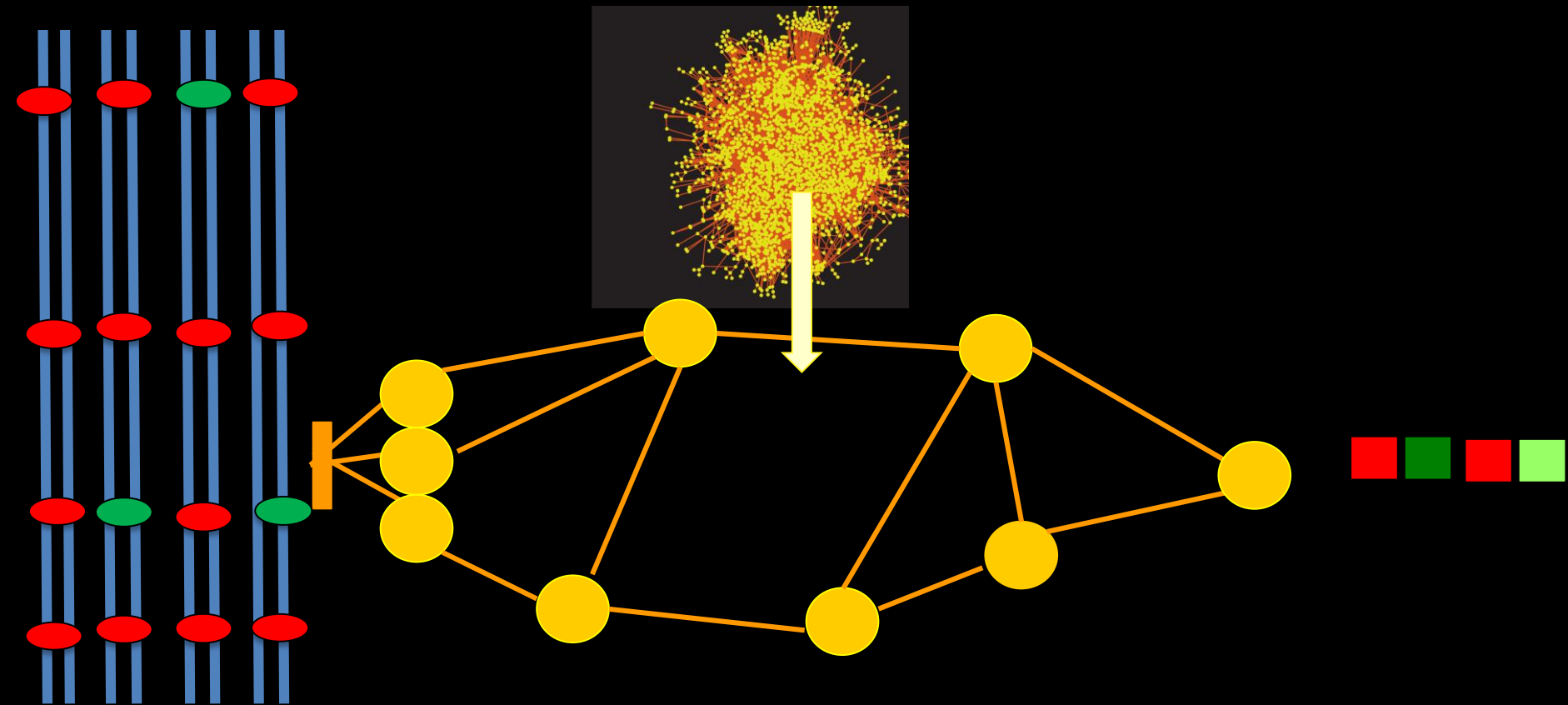
**Cancer Cases  
Gene expression data**

# eQTL analysis links expression variability to genotypic variability



Tu *et al* Bioinformatics 2006  
Suthram *et al* MSB 2008  
Kim *et al*. PLoS CB 2011/RECOMB 2010

# Uncovering pathways of information flow between CNV and target gene

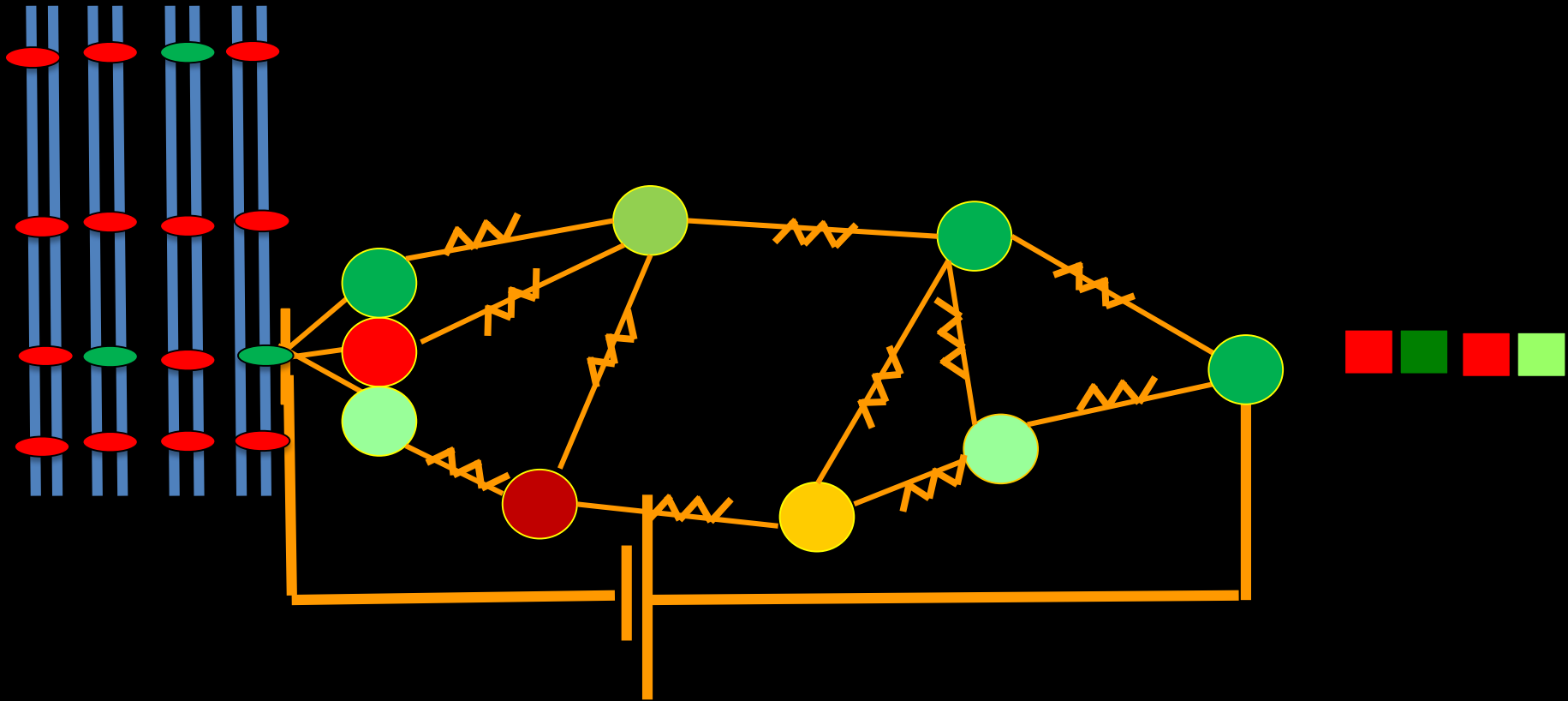


Tu *et al* Bioinformatics 2006

Suthram *et al* MSB 2008

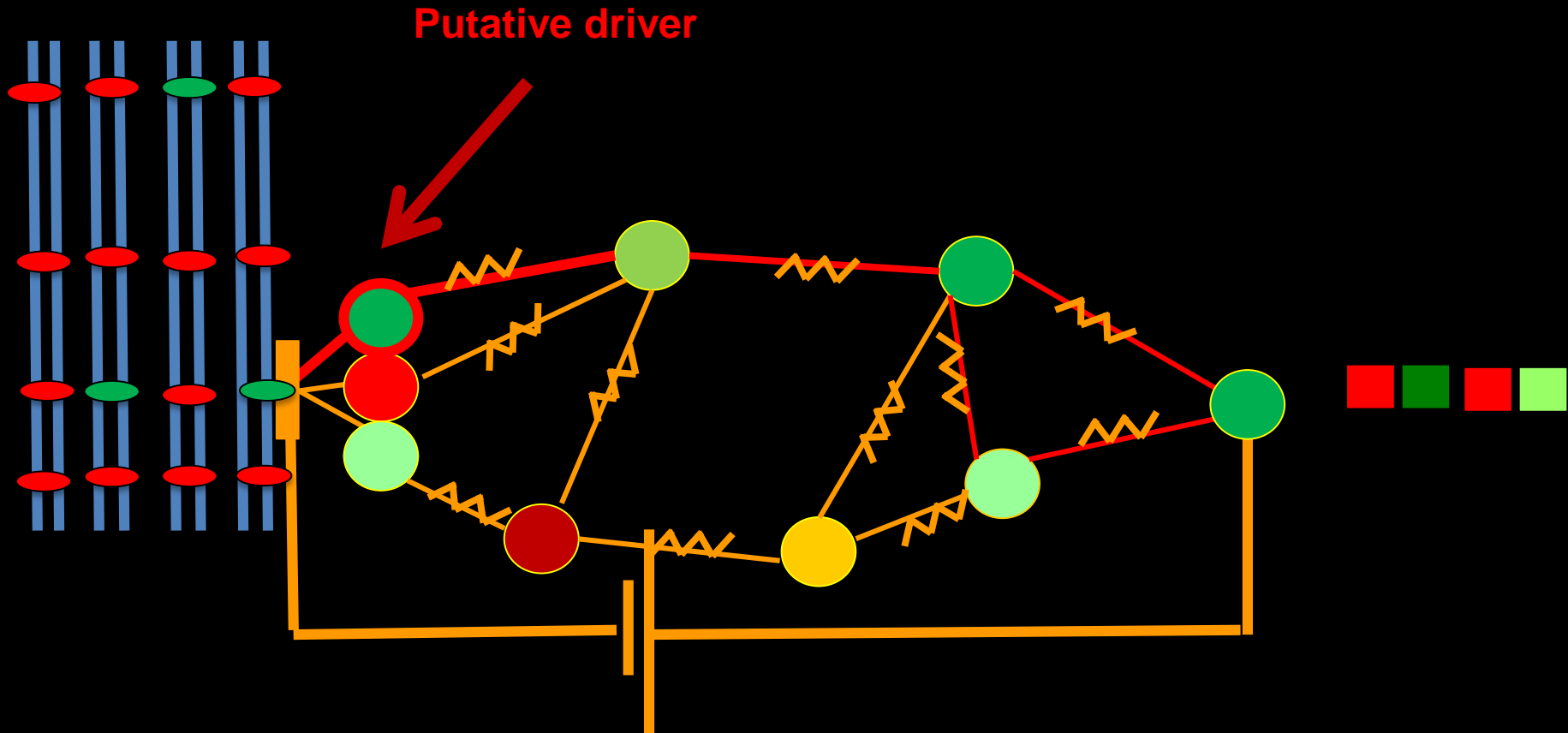
Kim *et al*. PLoS CB 2011/RECOMB 2010

# Adding resistances differentiate likelihoods of the edges

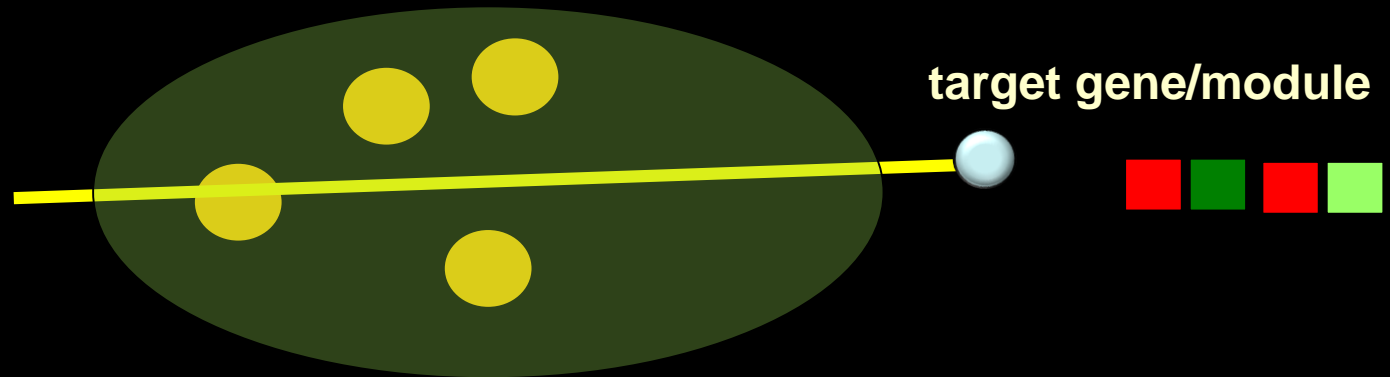
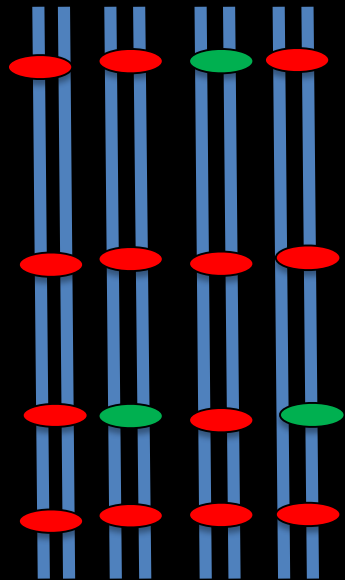


**Resistance** - set to favor most likely path -based on gene expression values  
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

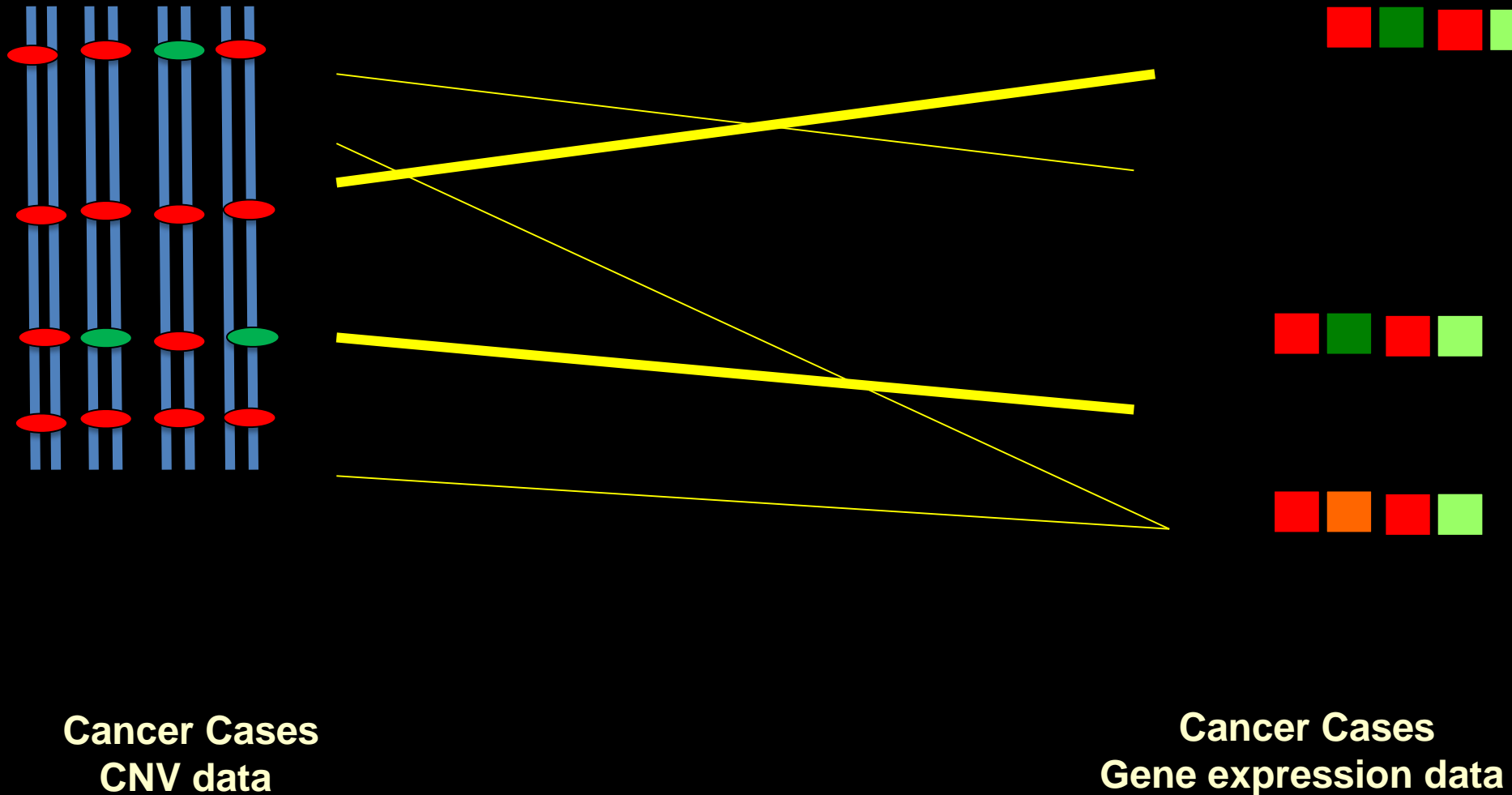
# Finding subnetworks with significant current flow



**Resistance** - set to favor most likely path -based on gene expression values  
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

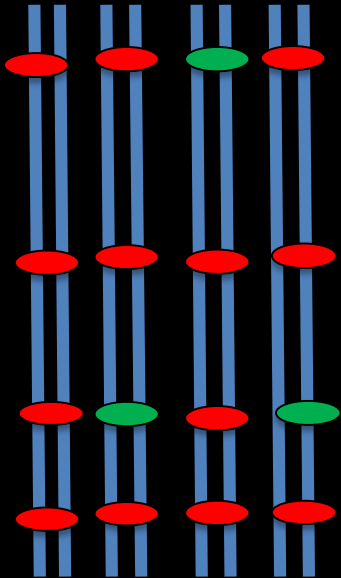


Repeat for other genes and significantly associated loci

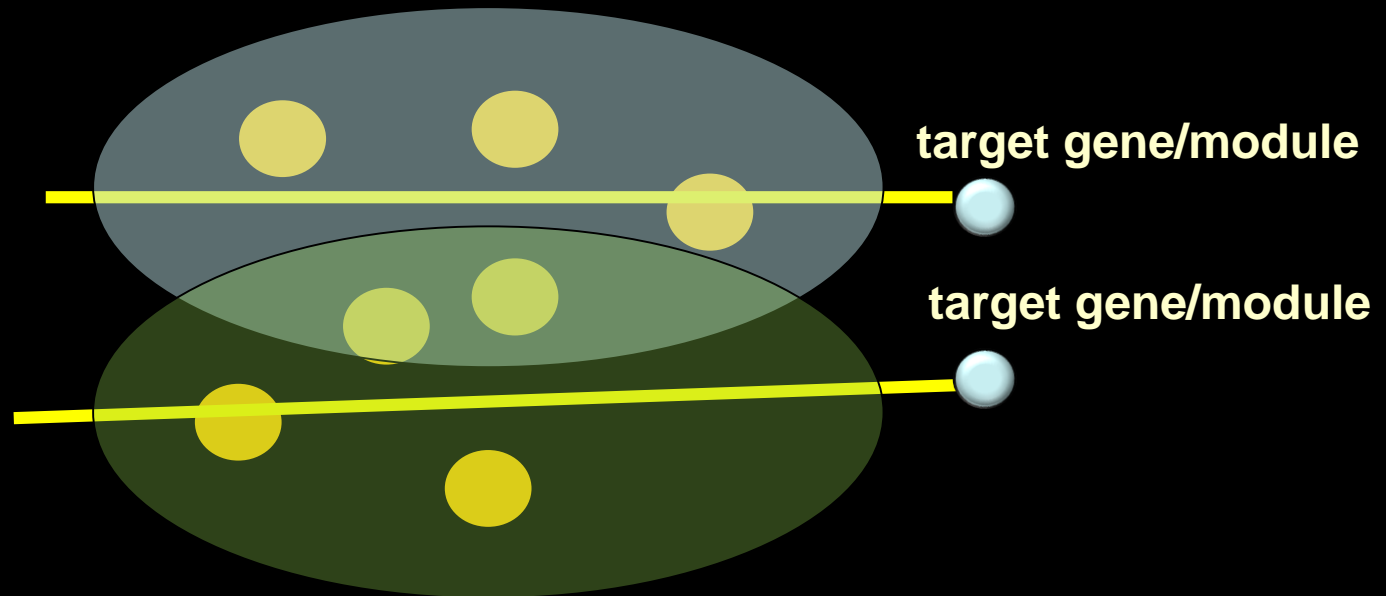


# Are there common functional pathways?

Cancer Cases  
CNV data



Cancer Cases  
Gene expression data

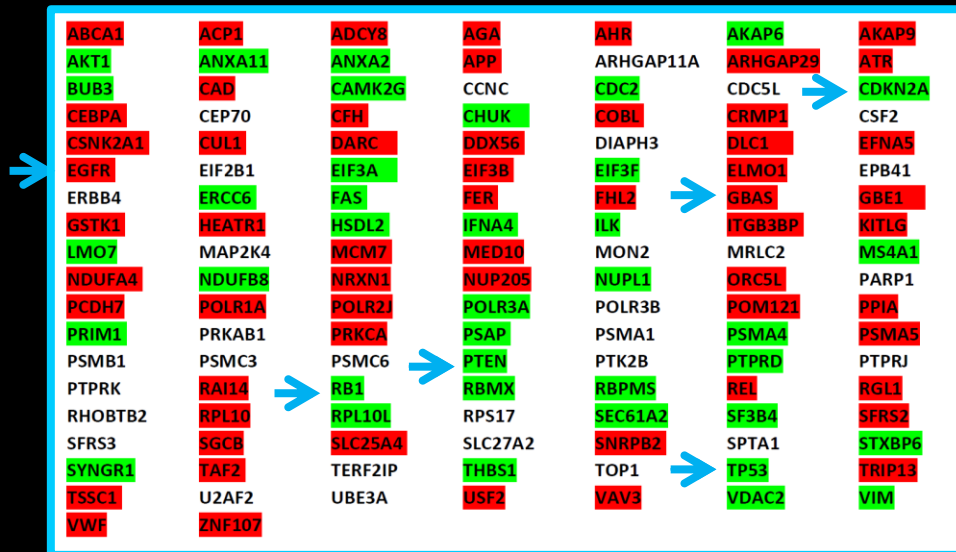


# Gene Hubs

MYC(110)	E2F1(88)	E2F4(43)	CREBBP(34)	GRB2(27)	SP3(26)	ESR1(25)
TFAP2A(25)	NFKB1(23)	MYB(22)	JUN(22)	E2F2(22)	RELA(21)	AR(21)
SP1(20)	RPS27A(20)	MAPK3(19)	POU5F1(17)	HIF1A(16)	PPARA(15)	CDC42(15)
UBA52(13)	CDK7(13)	YBX1(13)	YWHAZ(12)	CEBPB(12)	POU2F1(12)	UBE2I(11)
SMAD3(11)	TAL1(11)					

# Pathway Hubs

## Driving Copy number aberrations



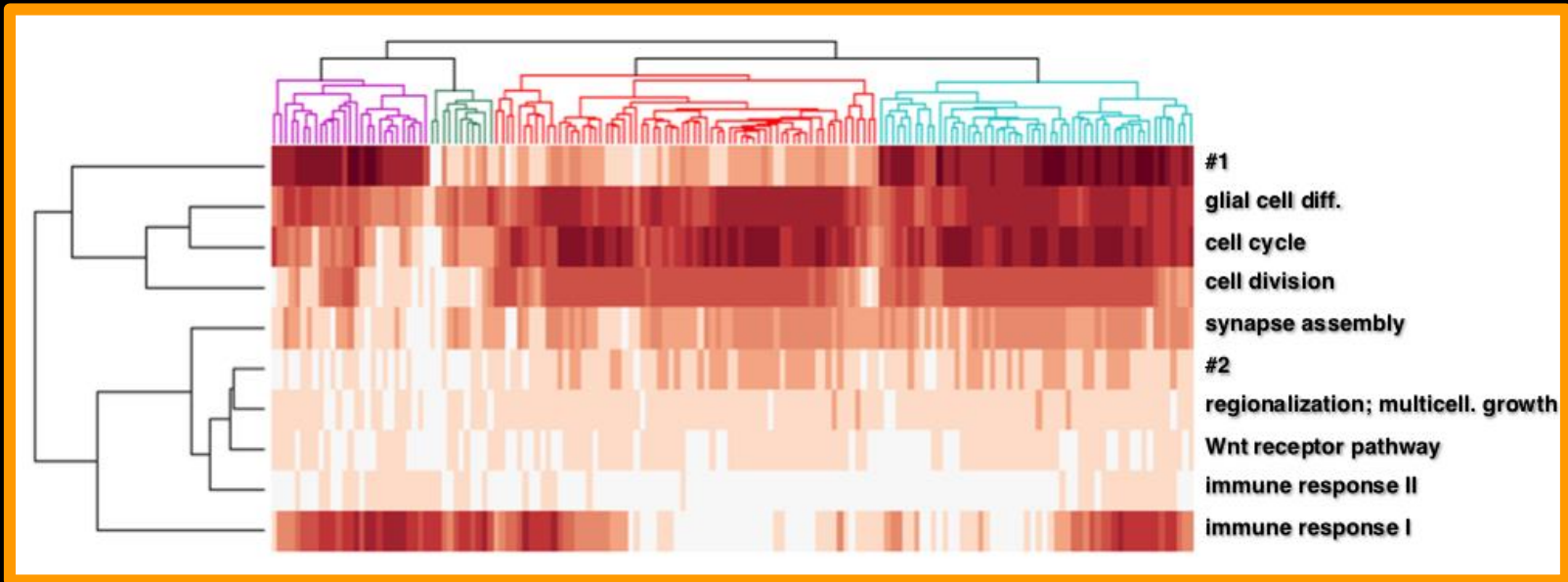
GO biological process	#
cell cycle arrest	10
epidermal growth factor receptor signaling pathway	9
negative regulation of cell growth	9
Ras protein signal transduction	9
regulation of sequestering of triglyceride	8
cell proliferation	7
nuclear mRNA splicing, via spliceosome	7
regulation of cholesterol storage	7
nucleotide-excision repair	7
RNA elongation from RNA polymerase II promoter	7
insulin receptor signaling pathway	6
transcription initiation from RNA polymerase II promoter	6
N-terminal peptidyl-lysine acetylation	5
phosphoinositide-mediated signaling	5
positive regulation of lipid storage	4
positive regulation of specific transcription from RNA polymerase II promoter	3
positive regulation of epithelial cell proliferation	3
base-excision repair	2
negative regulation of hydrolase activity	2
gland development	2
positive regulation of MAP kinase activity	2
regulation of nitric-oxide synthase activity	2
estrogen receptor signaling pathway	2
regulation of receptor biosynthetic process	2
response to organic substance	2
JAK-STAT cascade	2
regulation of transforming growth factor-beta2 production	2
G1/S transition of mitotic cell cycle	2
SMAD protein nuclear translocation	2

# Three general techniques that utilize network based approaches in cancer studies

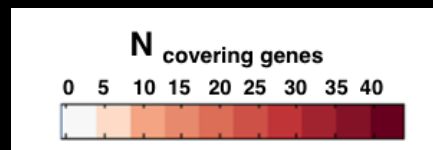
- Module cover
- Network Flow
- Mixture/topic models

# Different patients groups have different signature modules

cases



modules



# Phenotypic versus explanatory features

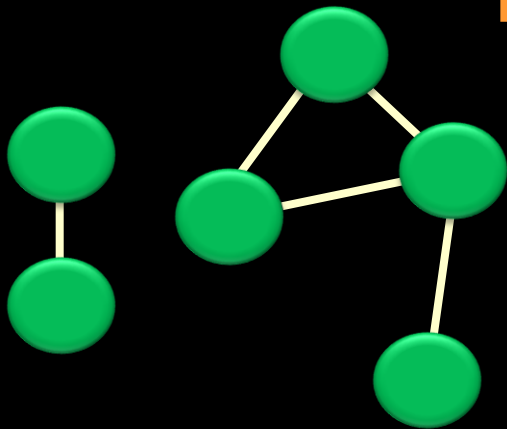
## Phenotypic features:

Survival time  
Response to drugs,.....  
Gene expression profile

## Explanatory features

- mutations, CNV, micro RNA level;
- Epigenetic factors,
- Sex, age, environment ....

## Patient graph



*Nodes* – patients

*Edges* – phenotypic similarities

## Key idea

neighbors in patient network should have similar explanatory features

# Assuming k subtypes, generate feature distribution for them

## Subtype I

EGFR\_A 0.45  
NF1\_M 0.37  
PTEN\_A 0.21

....

## Subtype II

PDGFA\_A 0.51  
IDH1\_M 0.29  
M53\_M 0.17

....

## Subtype III

mirR218\_H 0.38  
ICDK2\_D 0.22  
SHC1\_M 0.14

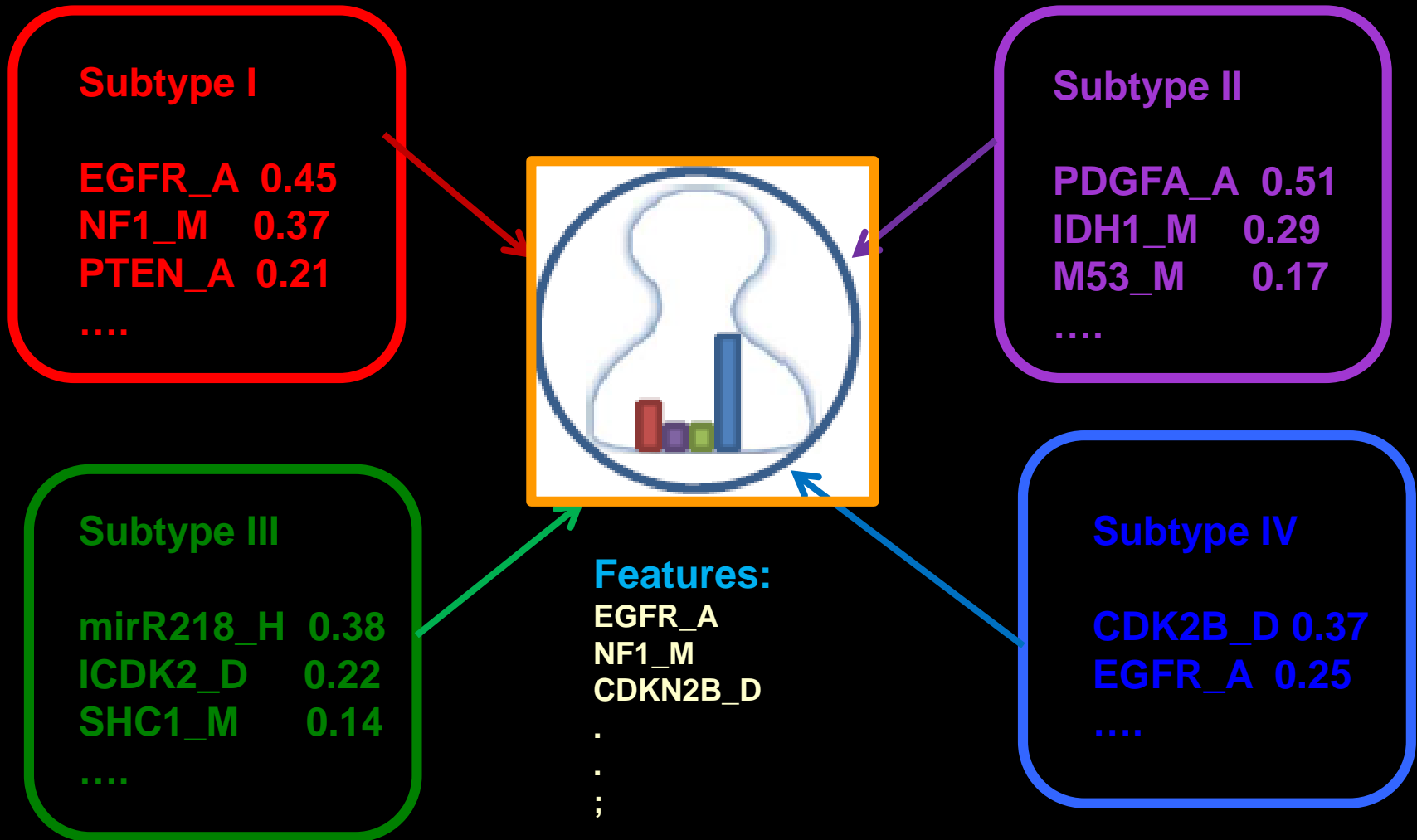
....

## Subtype IV

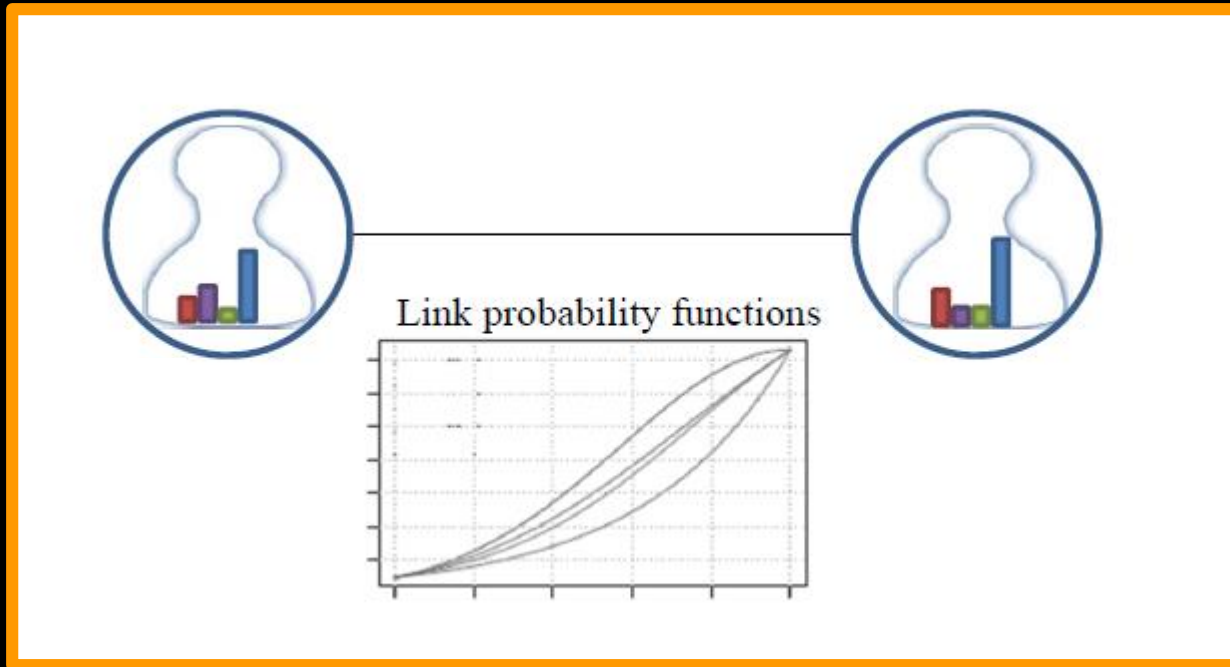
CDK2B\_D 0.37  
EGFR\_A 0.25

....

# Based on patient's features represent each patient as mixture of the subtypes

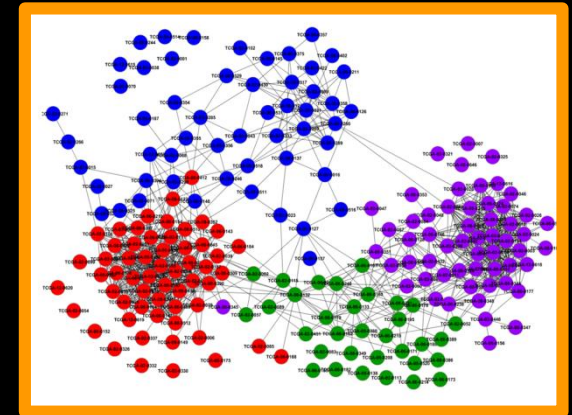
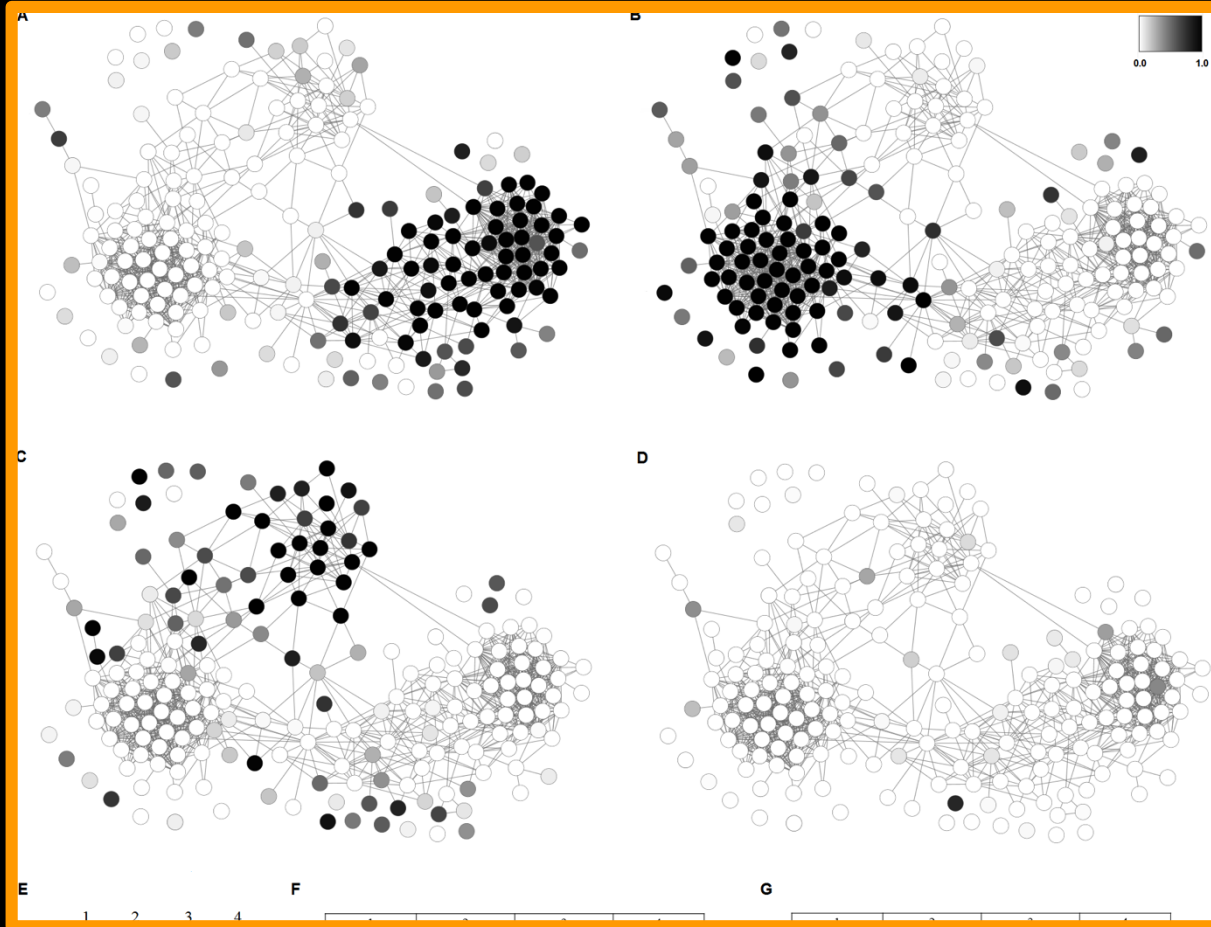


## Generate edges based on similarity of subtype mixtures

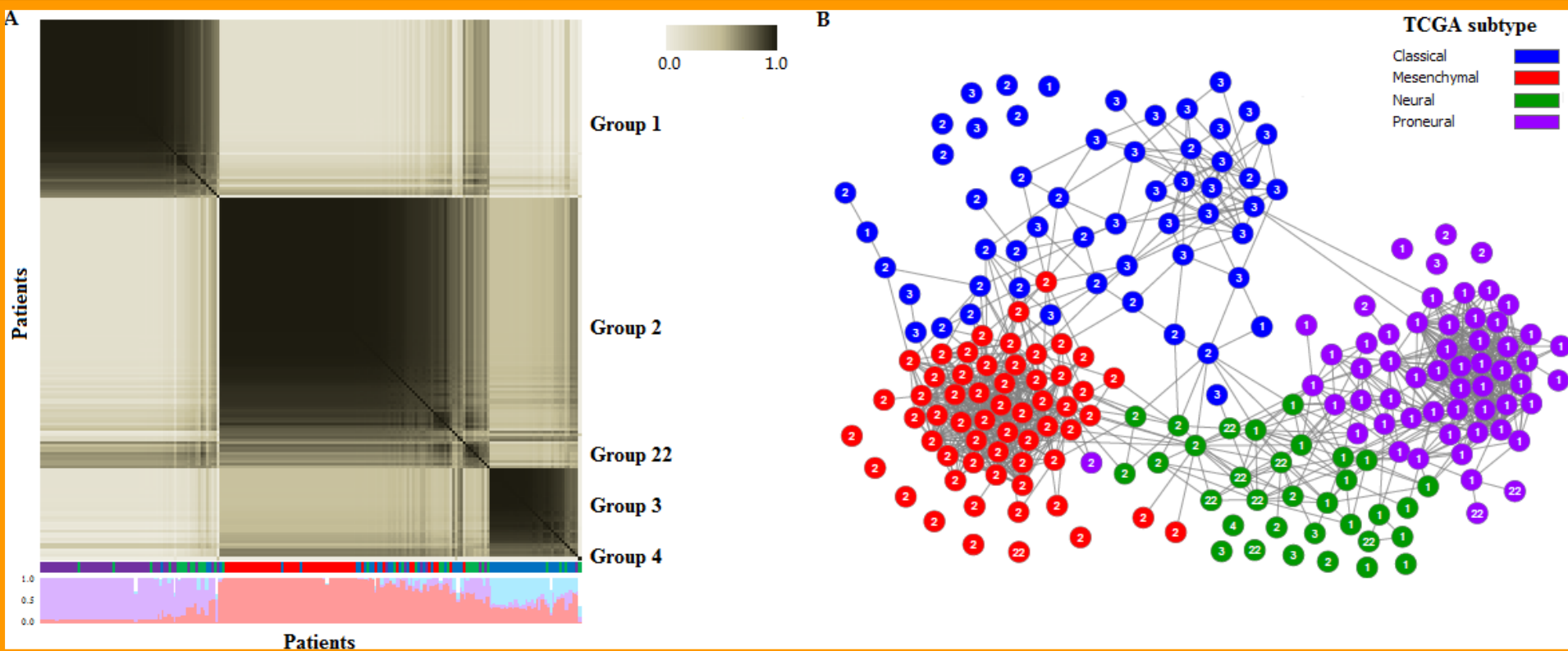


**Optimize parameters to maximize likelihood of the patient -patient network**

# Visualization of subtypes distribution form a sample model

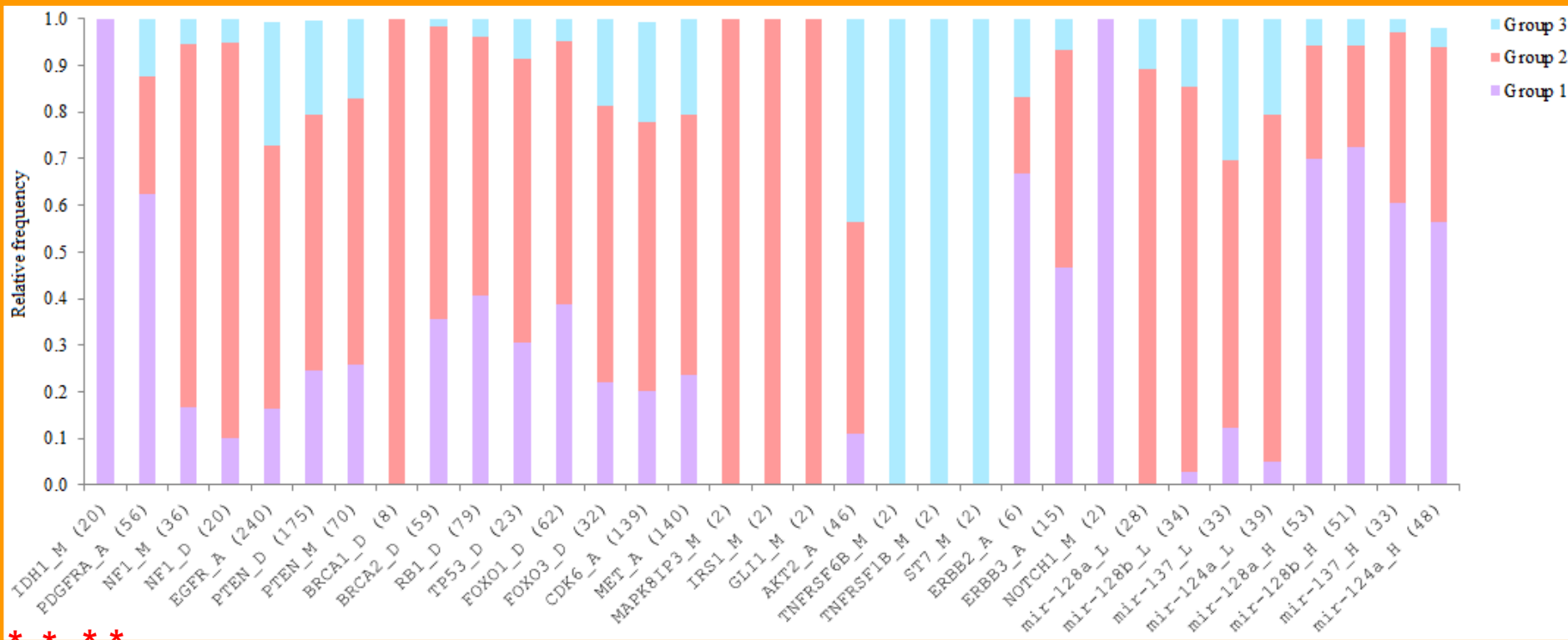


# Patient-patient relationship based on 1000 models



**Observation:** No separate Neural group

# Selected cancer related features



**Observations:** correctly recovered features from Varhaak et al. (TCGA)

**AKT2** – most important defining feature of the Classical group

**Potential benefits of using dys-regulated pathways as features**

## Summary

- Networks/Systems based approaches provide new view of cancer data
- These methods are general and can be adopted to new types of data

## Challenges

- Noisiness and incompleteness and bias of interactome
- More data is needed to be able to account for age/sex/environment and other complex dependencies

# Acknowledgments

## Przytycka's group

DongYeon Cho

YooAh Kim

Phuong Dao

Xiangjun Du

Damian Wojtowicz

Jan Hoinka



*Support: Intramural research program NLM / NIH*

# Using 1,000 models to infer:

- Probabilistic relation between patients
- Probabilistic relation between features
- Probabilistic relation between features and patients

## (Glioblastoma Multiforme)

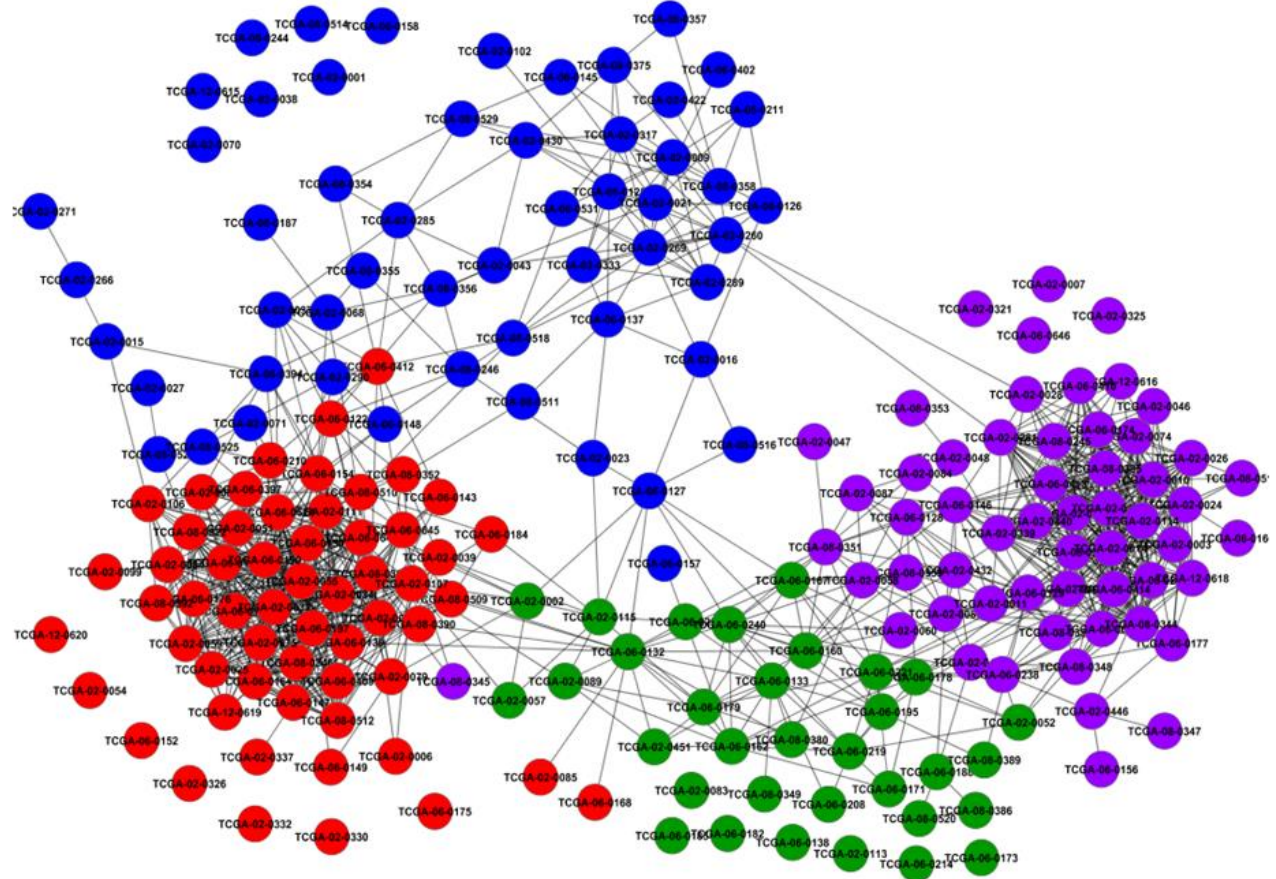
## Varhaak et al. Classification

## patient network for GMB

**Mesenchymal**

 **Classical**

## Proneural

 Neural

# Simultaneous modeling of phenotypic and explanatory features

In each model we assume

- $k$  subtypes
- each subtype is defined by probability distribution of (explanatory) features
- each patient is a mixture of these subtypes
- patients with similar phenotypic features have mixtures

# Visualization of subtypes distribution form a sample model

